

# Image Dataset Curation and Art Recommendation

**Antonio Rueda-Toicen**  
AI Engineer

**KI Service Zentrum, Hasso Plattner Institute**  
January 2024

**KI** Service  
Zentrum  
by Hasso-Plattner-Institut

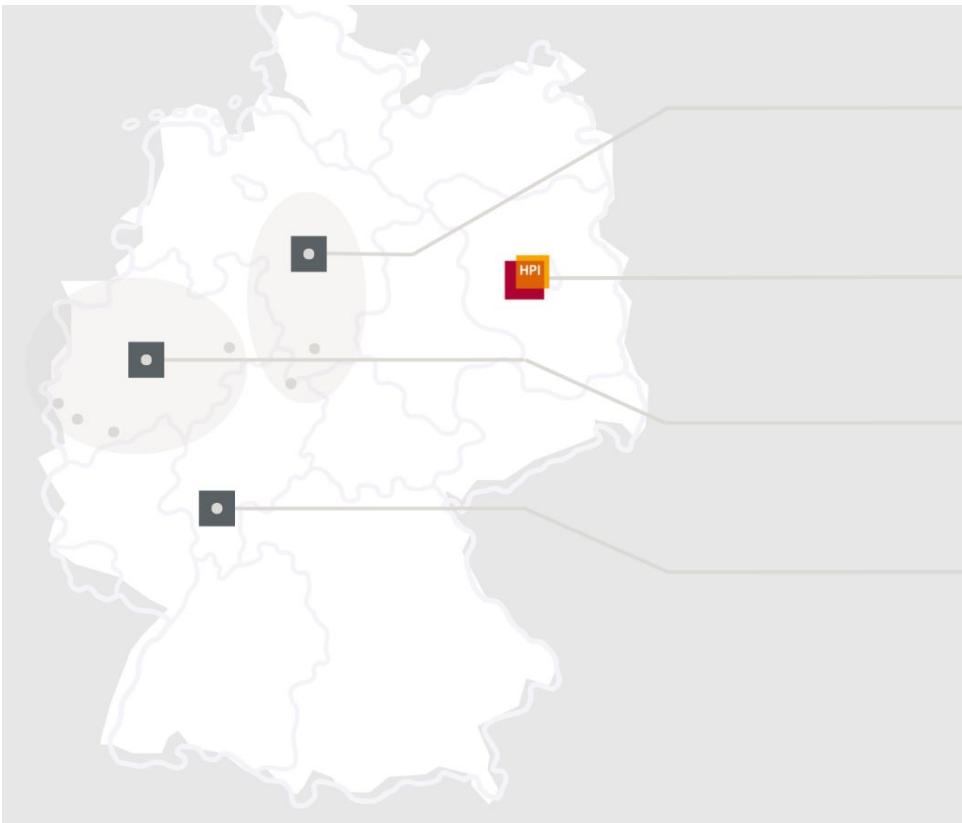


SPONSORED BY THE



Federal Ministry  
of Education  
and Research

<https://hpi.de/kisz/home.html>



## KISSKI

Hannover | Göttingen | Kassel  
Sensitive and critical infrastructures  
Focus: Medicine & Energy

## KISZ BB

Berlin | Brandenburg  
Educational and advisory services  
Use of AI in business and society

## West AI

Dortmund | Bonn | Jülich | Aachen | Paderborn  
Large and transferable AI models

## hessian AI

Service Center

Darmstadt  
Explainability, generalizability  
and contextual adaptation

<https://hpi.de/kisz/home.html>

# Agenda

- Brief intro and learning objectives
- Use cases for image similarity
- Understanding image embeddings
- Scraping images from Google Images or Bing
- Setting up Google Colab and Drive
- Using pre-trained networks for image similarity
- Visualizing embeddings with TensorBoard
- Review questions and discussion

# What we expect you to have

- Some Python knowledge
- Curiosity :)

# Learning objectives

At the end of the first workshop you will be able to:

- Describe use cases for image similarity in dataset curation
- Scrape images from Google Images or Bing
- Generate embeddings for images using a pretrained neural network
- Compare image pairs using cosine similarity
- Visualize embeddings in 3D using Tensorboard

# Learning objectives

At the end of the second workshop you will be able to:

- Visualize image neighborhoods with k-nn
- Cluster images using k-medoids
- Select representative images
- Classify images using a pre-trained Resnet
- Run zero-shot classification with CLIP

# How are we doing this workshop

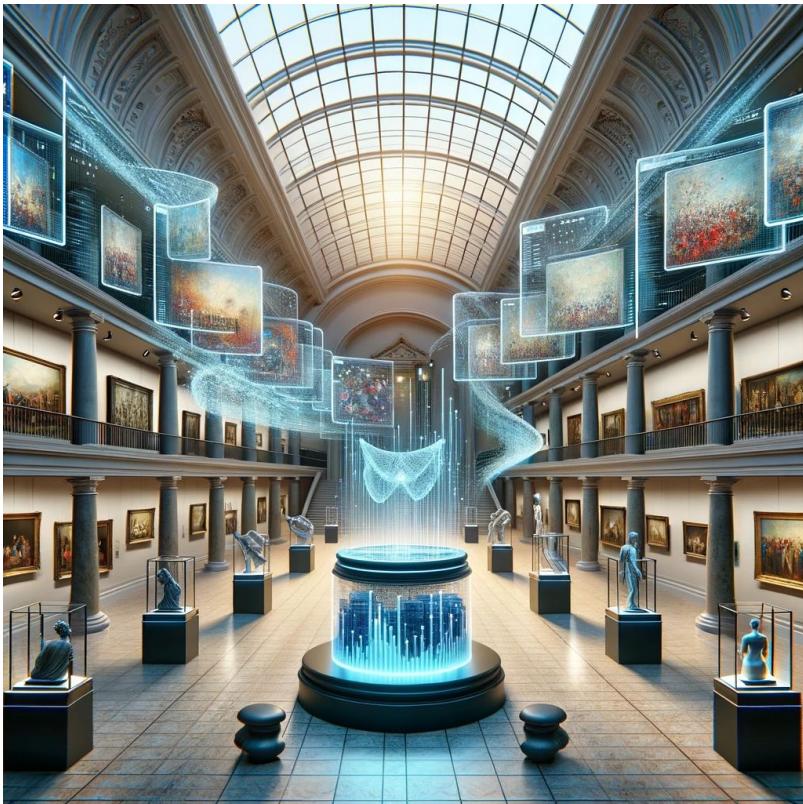
- We **type** most of the Python code from the tutorial notebooks in Google Colab

Repository:<https://github.com/KISZ-BB/image-dataset-curation-workshops>

# What you need

- A Google user account
- A Google Drive account with enough free space
- Google Chrome or Firefox

# What is dataset curation?



- We want to make data **accurate** and **relevant**
- We clean, deduplicate, and label
- This is similar to what museum curators do

# How similar are these two images?



Cosine similarity = 0.91

# How similar are these two images?



Cosine similarity = 0.78

# How similar are these two images?



Cosine similarity = 0.53

# How similar are these two images?



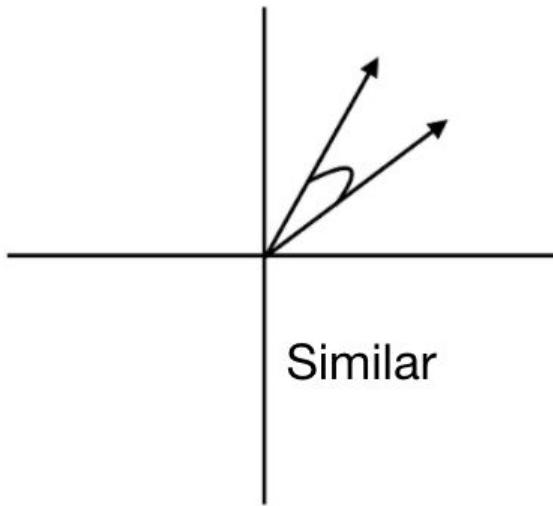
Cosine similarity = 0.65

# How similar are these two images?

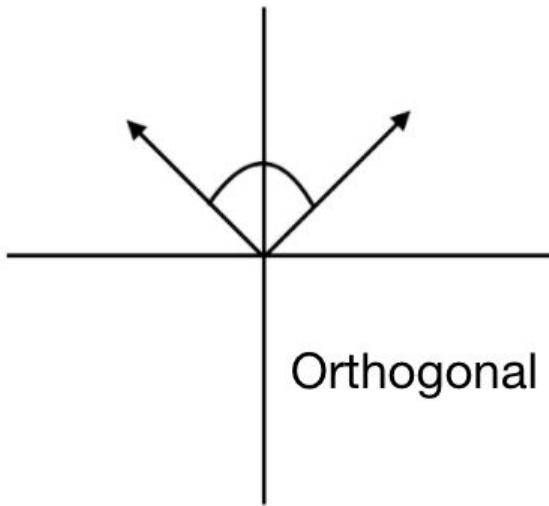


Cosine similarity = 0.84

# Cosine Similarity



Similar



Orthogonal

16

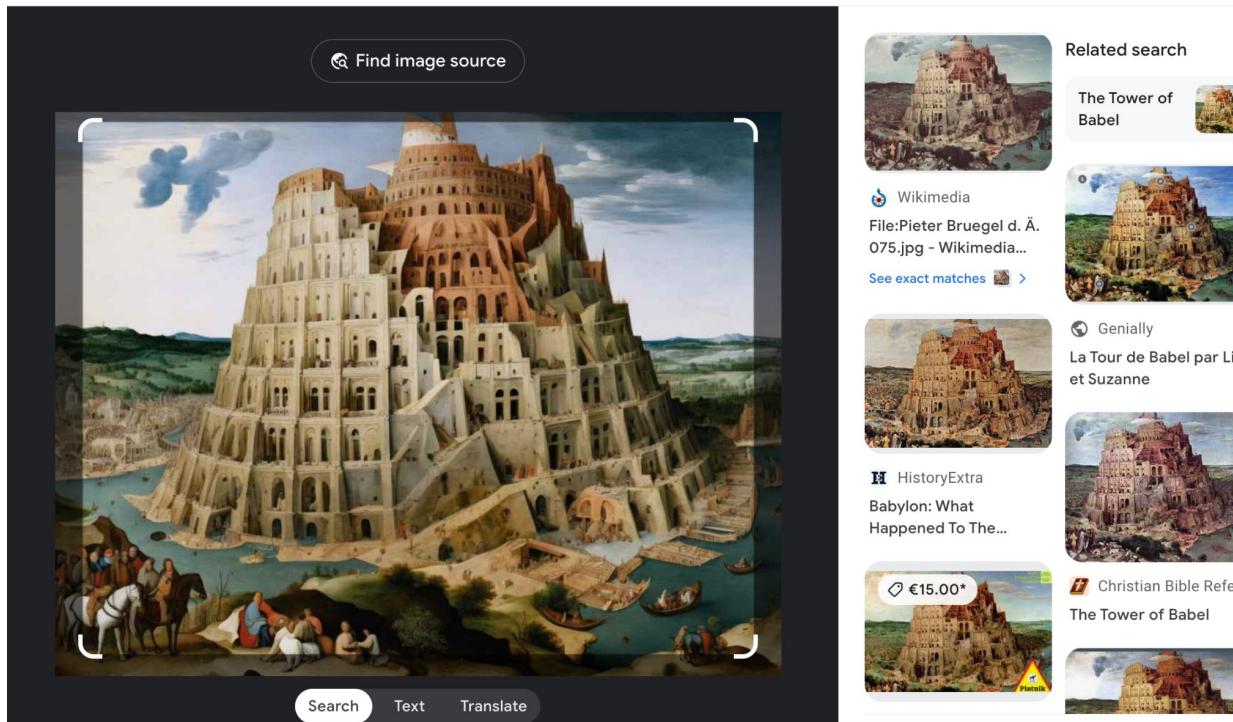
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Time for the first notebook

Comparing apples to apples

# Use Case: Google's reverse image search

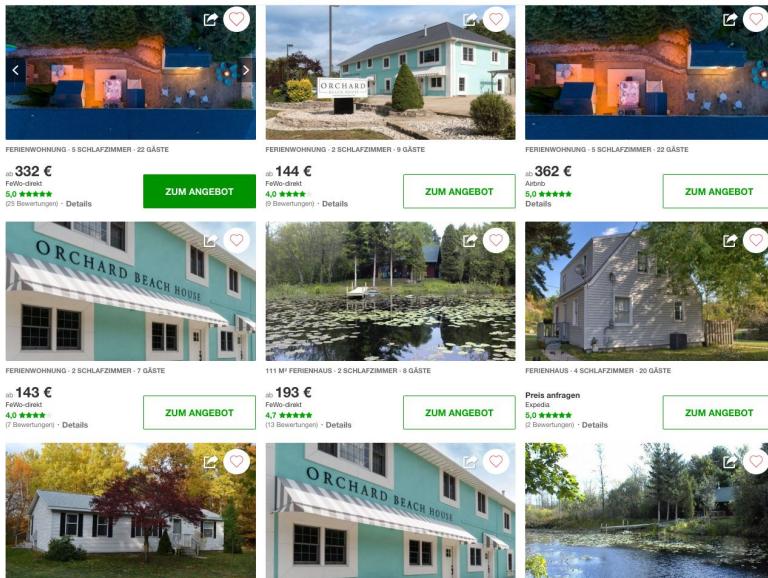
Google



<https://images.google.com/>

# Use case: image matching at hometogo.com

- Inventory understanding (500 million images)
- Providing the best deals to users (sample use case: strike prices)



3-STAR HOTEL · DOWNTOWN GRAND LAS VEGAS

\$30 ~~\$33~~  
Travelocity  
4.2 ★★★★★ · Details

[VIEW DEAL](#)

# Use case: primary image matching at Hometogo



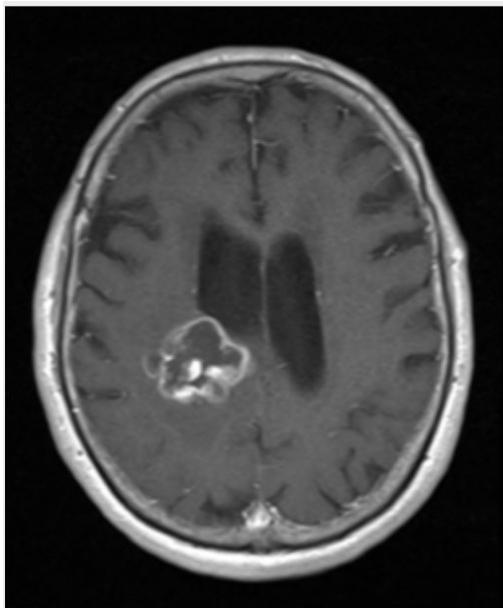
**Cosine similarity = 0.65**

# Use case: primary image matching at Hometogo

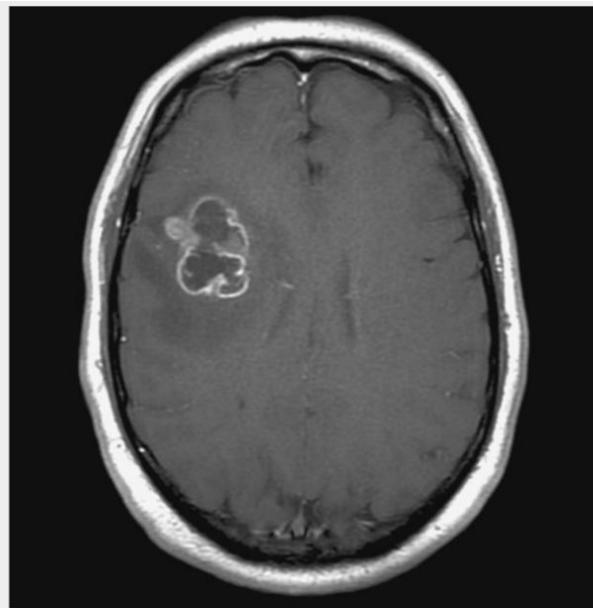


**Cosine similarity = 0.99**

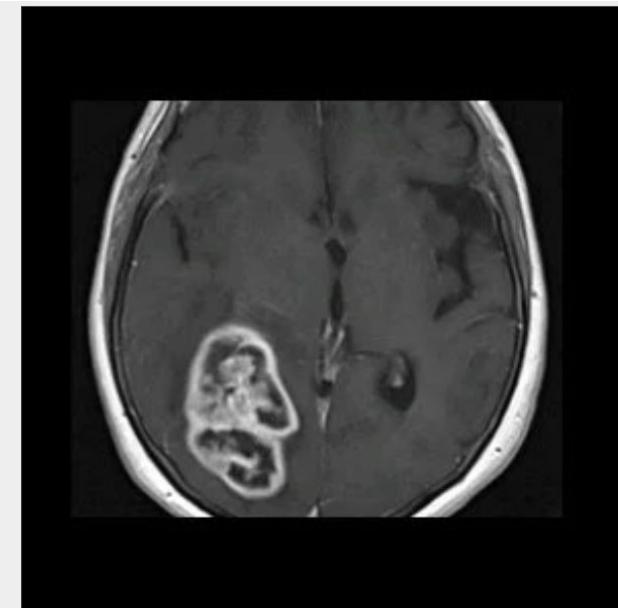
# Use case: differential diagnosis



Glioma grade I



Glioma grade I



Glioma grade III

# Use case: bike trip routing

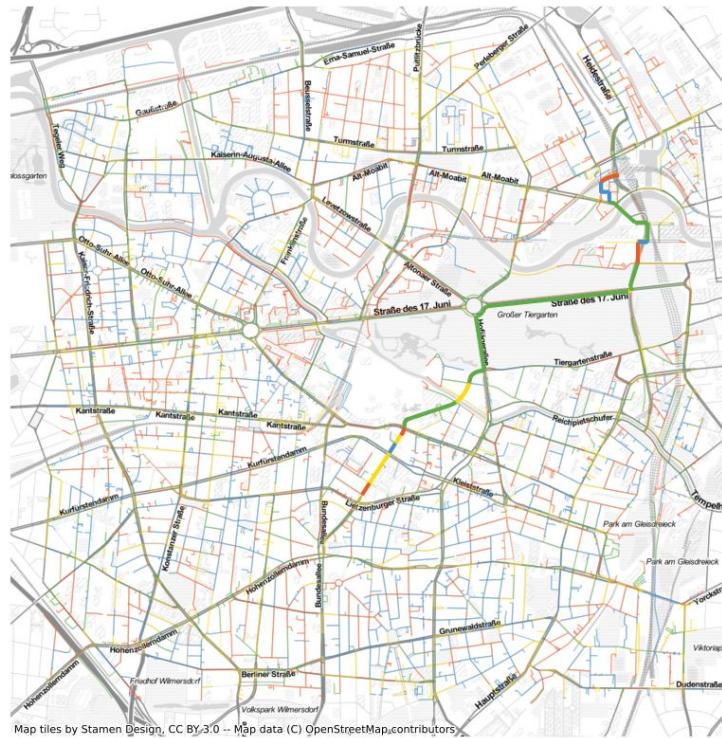
Cluster 0: Safe bike lanes



Cluster 1: Unsafe bike lanes



# Use case: bike trip routing



[www.github.com/graumannm/Berlin\\_Bike\\_CV](https://www.github.com/graumannm/Berlin_Bike_CV)

# Use case: cleaning image datasets



<https://github.com/cleanlab/cleanlab>

# Art Recommendation system

---

This is the repository of a portfolio project at DSR. This project aims to identify similar images using pre-trained computer vision networks. For an explanation of the technology see the [technology section](#).

## Contributors

---

- Catarina Ferreira
- Gargi Maheshwari

<https://github.com/gargimaheshwari/Wikiart-similar-art>

# Motivation: enhancing Wikiart's recommendations



The Painter and the Art Lover - Pieter Bruegel the Elder

## Pieter Bruegel the Elder

Pieter Brueghel de Oude

Born: c.1525; Breda, Netherlands (i)

Died: September 9, 1569; Brussels, Belgium (i)

Nationality: Flemish

Art Movement: Northern Renaissance

Painting School: Flemish School, Antwerp School

Genre: genre painting

Field: painting, printmaking

Influenced by: Hieronymus Bosch

Influenced on: Tobias Verhaecht, Peter Paul Rubens, Jan Miense Molenaer, Hendrick Avercamp

Art institution: Guild of Saint Luke

Friends and Co-workers: Maarten de Vos, Giulio Clovio

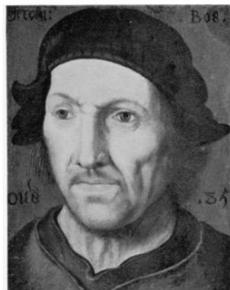
Family and Relatives: Jan Brueghel the Elder, Pieter Brueghel the Younger

Wikipedia: [en.wikipedia.org/wiki/Pieter\\_Bruegel\\_the\\_Elder](https://en.wikipedia.org/wiki/Pieter_Bruegel_the_Elder) ↗

<https://www.wikiart.org/en/pieter-bruegel-the-elder>

# Motivation: improving Wikiart's art recommendation

## RELATED ARTISTS i



**Hieronymus Bosch**

c.1450 - 1516



**Gregorio Lopes**

c.1490 - 1550



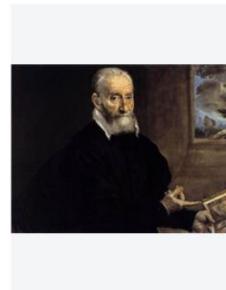
**Marinus van  
Reymerswaele**

c.1490 - c.1546



**Hans Holbein the  
Younger**

c.1497 - 1543



**Giulio Clovio**

1498 - 1578



**Jan van Hemessen**

c.1500 - c.1566



**Cristovao de  
Figueiredo**

c.1500 - c.1543

<https://www.wikiart.org/en/pieter-bruegel-the-elder>

# Motivation: improving Wikiart's art recommendation

## ARTISTS BY ART MOVEMENT

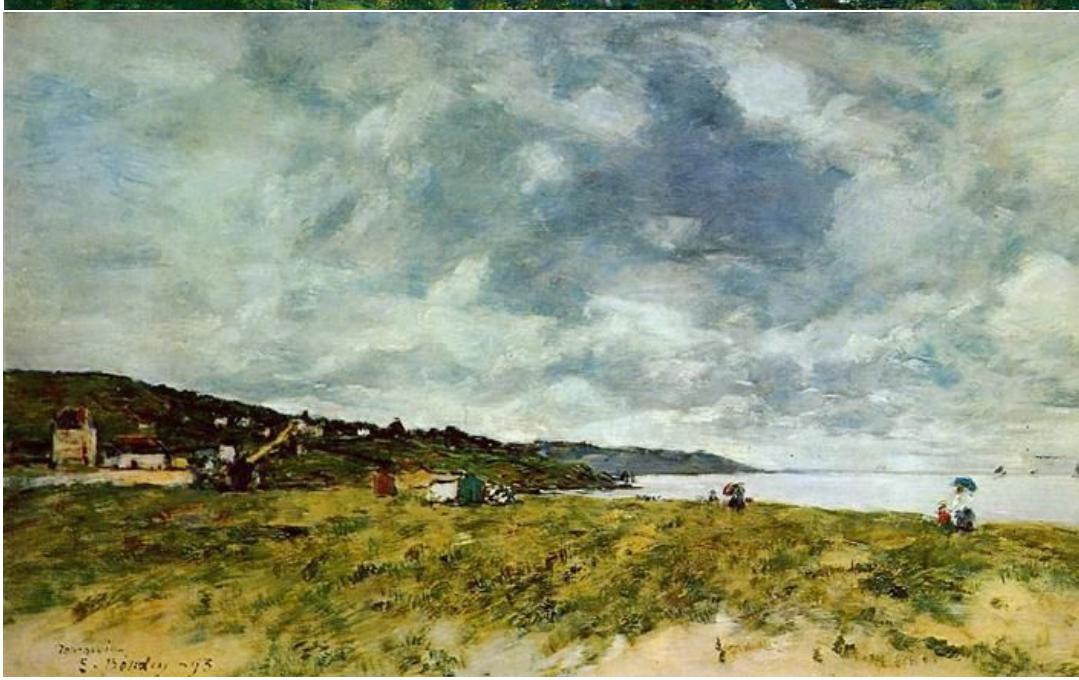
---

Art historians employ a number of ways to group world arts into systems of classification. They subdivide the continuous flow of artworks through time and space into groupings. These groupings are defined by the perception that the artworks within them share a single quality or a set of qualities that are significant. Significant qualities reflect a specific approach of an artist; they can include the formal, stylistic, iconographic, thematic, or other aspects of art. The definition of a grouping reflects judgments about the nature of meaningful connections between artworks, and between art and its larger context. Western arts are usually structured by art movements, using mostly cultural and aesthetic criteria, while Eastern arts are subdivided into periods according to political-dynastic markers.

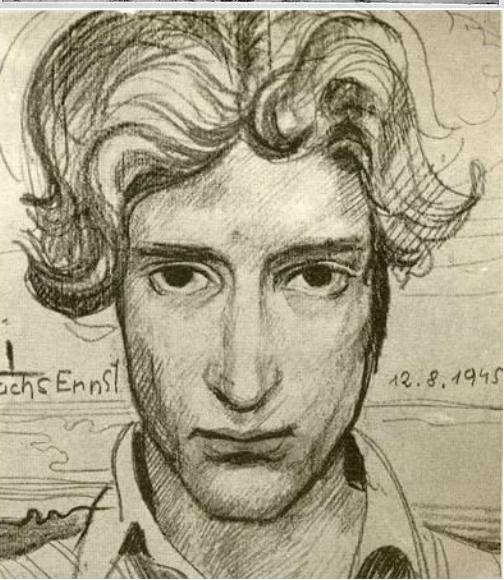
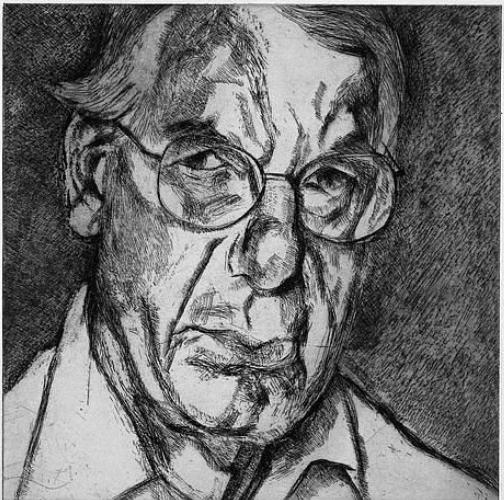
by time

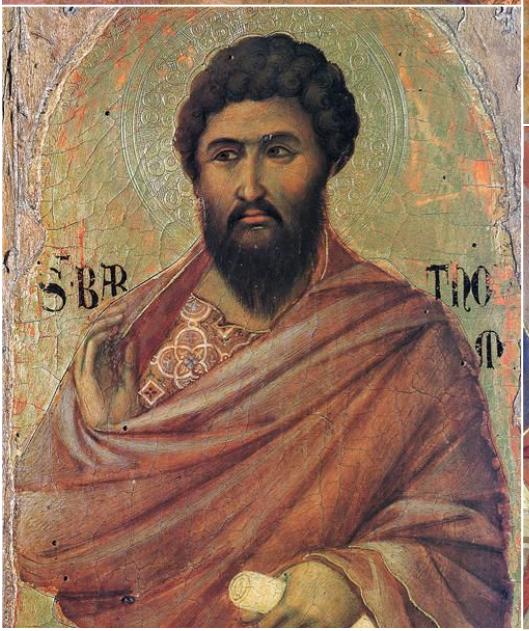
by name

by count

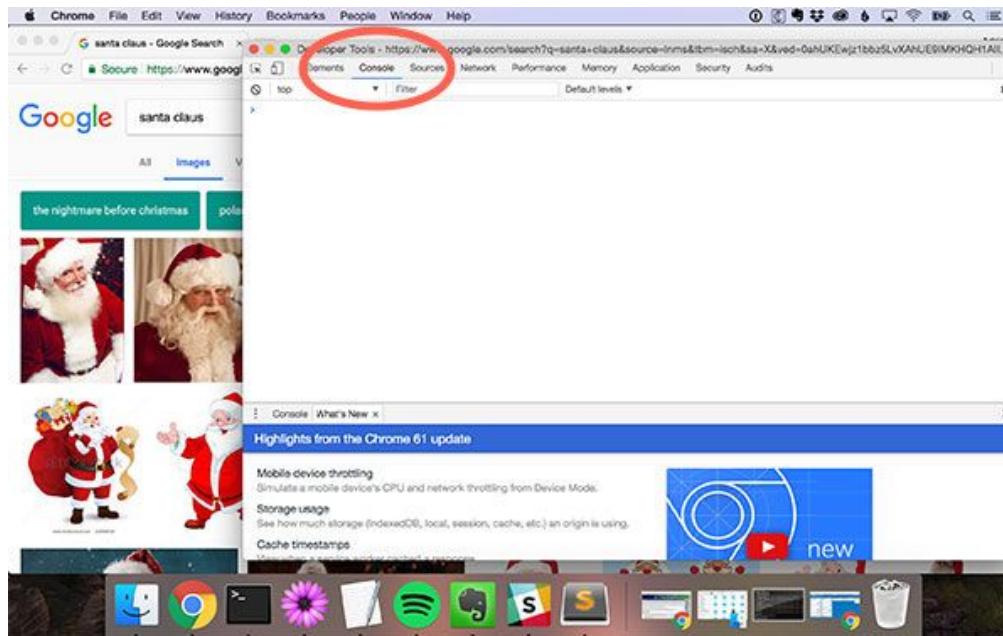






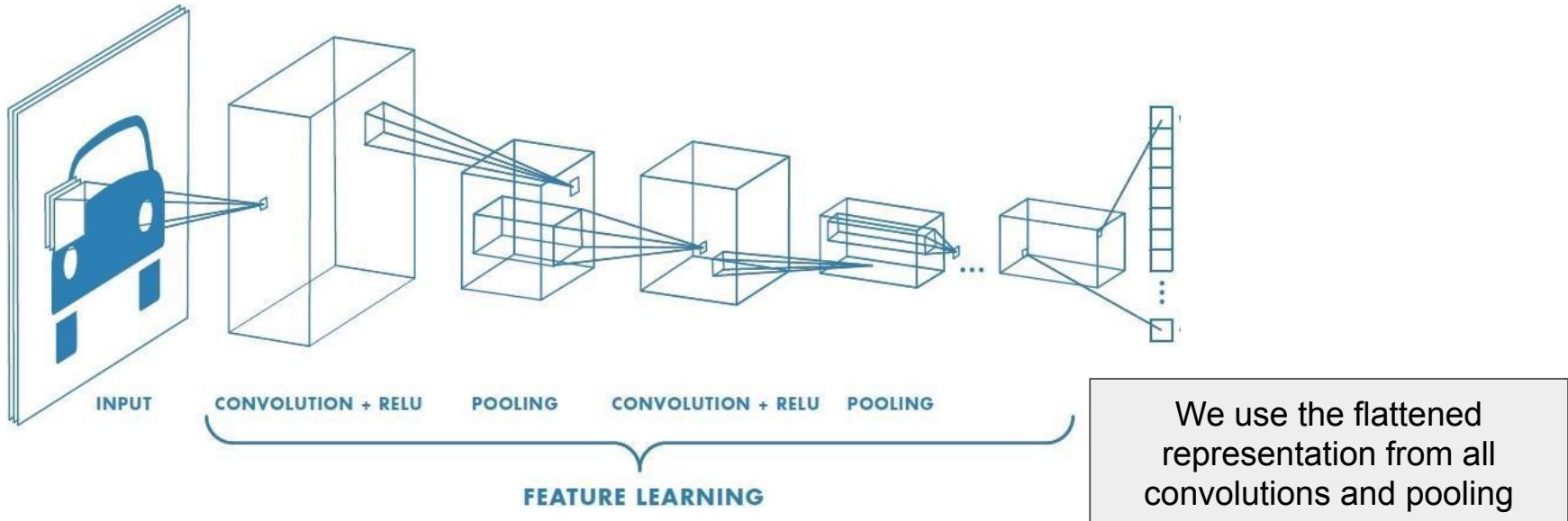


# Scraping a dataset from Google Images

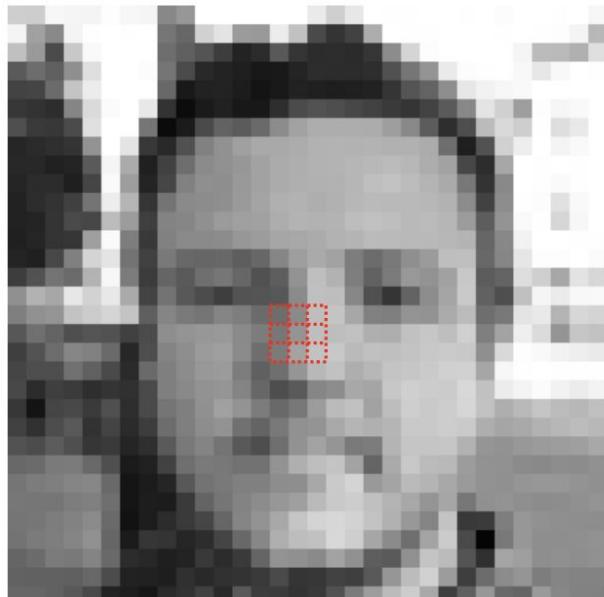


[PylImageSearch guide](#)

# Image Embeddings from Convolutional Networks



# Intuitions about convolutions



input image

$$\left( \begin{array}{ccc} 110 & + & 139 & + & 191 \\ \times 0 & & \times -1 & & \times 0 \\ \\ + & 120 & + & 149 & + & 191 \\ \times -1 & & \times 5 & & \times -1 \\ \\ + & 124 & + & 164 & + & 195 \\ \times 0 & & \times -1 & & \times 0 \\ \\ = & 131 \end{array} \right)$$

kernel:



output image

<https://setosa.io/ev/image-kernels/>

# The ImageNet dataset



14,197,122 images, 21841 synsets indexed

[Home](#) [Download](#) [Challenges](#) [About](#)

Not logged in. [Login](#) | [Signup](#)

**ImageNet** is an image database organized according to the **WordNet** hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. The project has been **instrumental** in advancing computer vision and deep learning research. The data is available for free to researchers for non-commercial use.

<https://www.image-net.org/>

# Visualizing embeddings in TensorBoard



# Image neighborhoods

1.0



0.73



0.73



0.72



0.7



0.7



# Review questions

- How can we access an image array from its file's url?
- What is an embedding?
- What is maximum possible value of cosine similarity between two vectors?
- What's the cosine similarity between two orthogonal vectors?
- What is Imagenet?

# Join us for the next workshops!

- [AI Service Center - Berlin Brandenburg](#)

## Next topics:

- Image clustering
- Fine tuning models for image classification
- Image analysis with CLIP
- Meta's Segment Anything
- Segmentation and object detection with Detectron2
- Using Qdrant as a vector database for images
- Model deployment with FastAPI and Docker